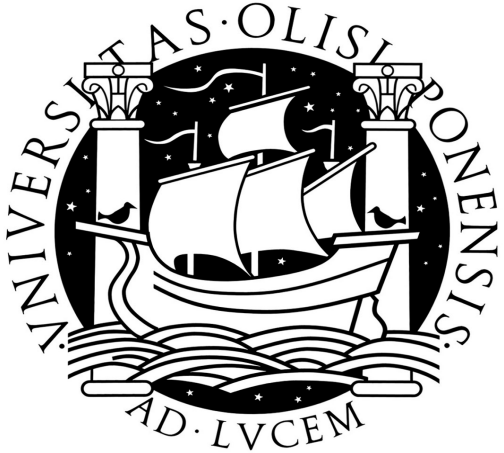


Grease-II Workshop

Ontology-driven GeoSimilarity

Francisco Couto

Universidade de Lisboa, Faculdade de Ciências,
Departamento de Informática

fcouto@di.fc.ul.pt

What are we doing?

- **Task 4: Geographic Information Retrieval: geographicity, geographic pseudo-relevance feedback, geo-similarity**
  - ***Finally, this task is also concerned with the study of similarity measures for geographic terms for geographic retrieval.***
  - ***…heuristics for geographical similarity, such as:***
    - ***spatial proximity, topological containment, semantic similarity in the geographic ontology, and importance of geographic concepts***

Why?

- *...computing the geographic scope of a document, selecting the place(s) that best cover each document through the use of the ontology, as an alternative to using all the extracted references that were extracted*

Why?

- **Task 3:Geographic Information Extraction and Knowledge Integration**
  - *... we need to research statistical methods for validating the extracted data similar to those we have been developing in the past years for mining the biological literature.*
    - CAC: validating by comparing extracted data with manual curated one

What is it?

- When entities are described using a common schema they can be compared by means of their annotations.
  - This type of comparison is called semantic similarity, since it assesses the degree of relationship of two entities by the similarity in meaning of their annotations.

Approaches

- Several approaches are possible
  - to quantify semantic similarity
  - between terms or annotated entities in an ontology
  - represented as a directed acyclic graph (DAG)

Scope

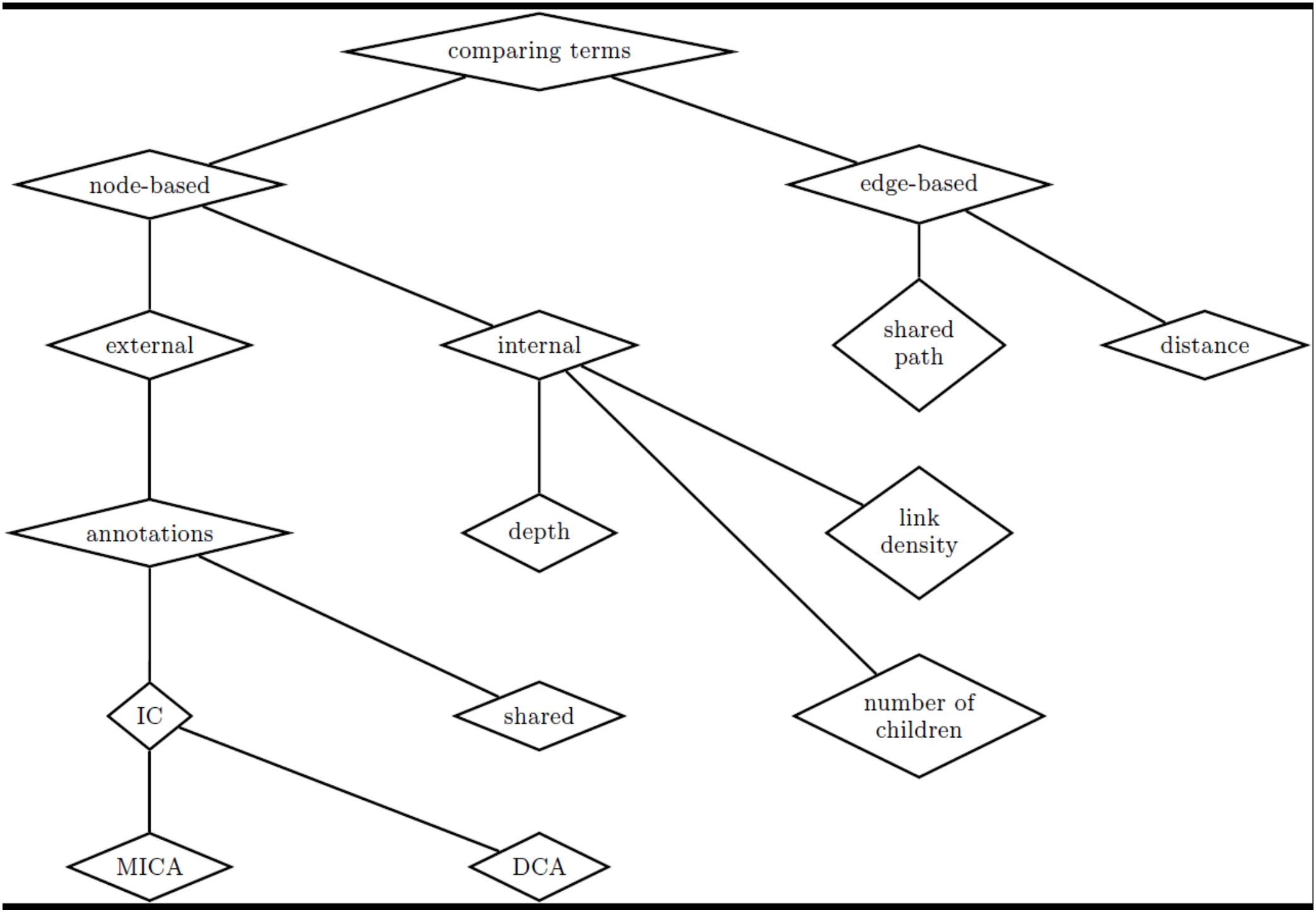
- which concepts to compare:
  - geoentities;
  - or web-pages annotated with geoentities;

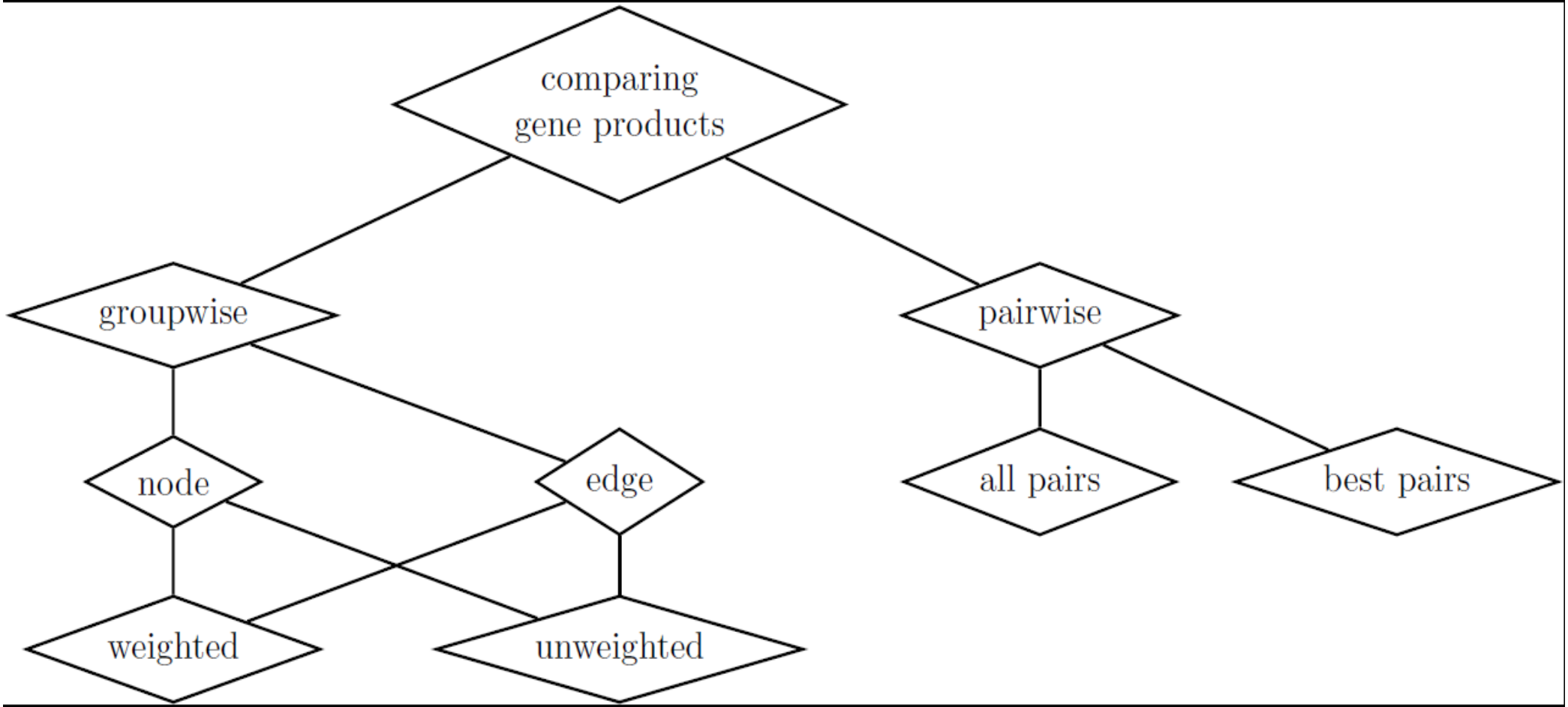
Data Source

- which components of the ontology they use, i.e.
  - edges vs. nodes;

Measure

- how they quantify and combine the information stored on those components.


comparing terms
node-based
edge-based
external
internal
shared path
distance
annotations
depth
link density
IC
shared
number of children
MICA
DCA


comparing
gene products
groupwise
pairwise
node
edge
all pairs
best pairs
weighted
unweighted

Measure	Approach	Techniques
Resnik	node-based	MICA
Lin	node-based	MICA
Jiang and Conrath	node-based	MICA
GraSM	node-based	DCA
Schlicker et al.	node-based	MICA
Wu et al. (2005)	edge-based	shared path
Wu et al. (2006)	edge-based	shared path; distance
Bodenreider et al.	node-based	shared annotations
Othman et al.	hybrid	IC/depth/number of children; distance
Wang et al.	hybrid	shared ancestors
Riensche et al.	node-based	IC/MICA; shared annotations

Measure	Approach	Techniques	Term Comparison
Lord et al.	pairwise	all pairs average	Resnik/Lin/Jiang
Sevillla et al.	pairwise	all pairs maximum	Resnik/Lin/ Jiang
Schlicker et al.	pairwise	best pairs	simRel
Azuaje et al.	pairwise	best pairs	Resnik/Lin/Jiang
Couto et al.	pairwise	best pairs	GraSM + (Resnik/Lin/Jiang)
Wang et al.	pairwise	best pairs	Wang
Posse et al. (XOA)	pairwise	all pairs maximum	XOA
Gentlman (simLP)	groupwise	edge-based unweighted	none
Gentlman (simUI)	groupwise	edge-based unweighted	none
Pesquita et al. (simGIC)	groupwise	node-based weighted/IC	none
Chabalier et al.	groupwise	node-based weighted/ IC	none

Node-based approaches rely on comparing the properties of the terms involved. The properties can be related to the terms *themselves* or their *ancestors* or their *descendents*. One concept commonly used in these approaches, is the information content (IC), which gives a measure how specific and informative a term is. The IC of a term c can be quantified as the negative log likelihood,

$$-\log p(c)$$

where $p(c)$ is the probability of occurrence of c in a specific corpus (e.g. UniProt knowledgebase). The probability of occurrence is normally estimated by the frequency of annotation of the node. The concept of IC can be applied to the common ancestors two terms have, to quantify the amount of information they share, and thus measure their semantic similarity. There are two main approaches for doing this: the most informative common ancestor (**MICA technique**), where only the common ancestor with the highest IC is considered [6] and the disjoint common ancestor (**DCA technique**), where all disjoint common ancestors are considered [7]. Alternatively, information content can also be calculated from the *number of children* a term has in the GO structure [8].

Results so far

- Information Content for each Geo-PT term
  - Calculate ancestors (done)
  - Annotations for each term (in progress)
    - Based on Google n-grams
- Problems:
  - Daniel had to study ontologies,
  - and understand the semantic similarity applications

Working Plan

- API (Java) for calculating semantic similarity in Geo-PT
  - Make it available as a web service
- Assessment
  - GEO-CLEF
  - validation of submitted results
- Integrate in GEO-Tumba
  - Validate extracted results
  - Improve the identification of geographic scope
  - Improve query disambiguation
  - …